Assignment 02

PUBH 8878: Heritability, Segregation, and Gene Mapping Theory

Due: [Date TBD]

Instructions

This assignment covers theoretical foundations of heritability analysis, segregation modeling, and gene mapping strategy. You will work with mathematical derivations, implement computational algorithms, and simulate genetic data to understand the statistical genetics toolkit.

Requirements: - Show all mathematical work - Submit well-documented R code with clear comments - Interpret results in biological context - Submit both `.qmd` source and rendered output

---

1. Recall (Young, 2022): "The first challenge is one of precision. The information used to estimate heritability from rare variants by GREML-WGS comes from the variation in sharing of rare variants among distantly related pairs of individuals [13, 15]. However, distantly related individuals typically do not share any particular rare variant, so the variation in rare variant sharing is low. This means that large samples with high quality WGS data are required to obtain precise estimates, and such samples are not common yet. Based on the only existing application of GREML-WGS [13], a sample size of ~40,000 would produce estimates precise enough to be statistically distinguished from other heritability estimates (Table 1). It is likely that this challenge will be overcome shortly, since samples of similar magnitude already exist [16]."

a. Assume the probability that two distantly related individuals share a rare variant is $p = 0.001$. What sample size is required to have a 95% chance that at least one pair shares a rare variant?

b. Let's say the true heritability of a trait is $h^2 = .4$. How many samples are required to achieve power. 80% to detect this heritability at $\alpha = 0.05$ using GREML? Use the `GCTA` software power calculator (https://yanglab.westlake.edu.cn/software/gcta/#Power%20Calculator) and assume the following parameters: 1) 1,000,000 SNPs; 2) prevalence = 0.1; 3) case-control ratio = 1:4; 4) disease risk = 0.01.

Problem 1: Recurrence Risk and Heritability Theory (25 points)

Part A: Mathematical Derivations (15 points)

Consider a binary trait with population prevalence $K = 0.01$ and sibling recurrence risk ratio $\lambda_{\text{sibling}} = 25$.

i. Using the liability threshold model, derive the relationship between λ_{sibling} and heritability h^2 on the liability scale. Start from the bivariate normal distribution of sibling liabilities and show that:

$$\lambda_{\text{sibling}} = \frac{\Phi_2(T, T; h^2/2)}{K^2}$$

where Φ_2 is the bivariate normal CDF and T is the liability threshold.

ii. For low prevalence traits ($K < 0.05$), prove that this simplifies to:

$$h^2 \approx \frac{\ln(\lambda_{\text{sibling}}) \cdot K^2}{i^2}$$

where $i = \phi(z_K)/K$ is the intensity of selection. Show all approximation steps.

iii. Calculate h^2 for the given parameters and compare to the exact formula.

Part B: Computational Implementation (10 points)

i. Implement the liability threshold model in R:

```r
# Template - you need to complete this function
liability_heritability <- function(lambda_sibling, K) {
    # Compute threshold
    z_K <- qnorm(1 - K)

    # Compute intensity of selection
    i <- dnorm(z_K) / K
    # YOUR CODE HERE: Implement both exact and approximate formulas
    # Return list with h2_exact, h2_approx, and comparison
}
```

Test your function with various $(\lambda_{\text{sibling}}, K)$ combinations and create a plot showing when the approximation breaks down.

Problem 2: Heritability Estimation and Bias (30 points)

Part A: Theoretical Analysis (15 points)

2.1 Consider the additive genetic model for a quantitative trait:

$$Y = \mu + \sum_{m=1}^{M} a_m X_m + \epsilon$$

where X_m is the number of effect alleles at locus m with allele frequency p_m.

(a) Prove that the additive genetic variance is:

$$\sigma_A^2 = \sum_{m=1}^{M} a_m^2 \cdot 2p_m(1 - p_m)$$

(b) Show that narrow-sense heritability $h^2 = \sigma_A^2/(\sigma_A^2 + \sigma_E^2)$ can be estimated from parent-offspring regression as the slope β_1 in:

$$Y_{\text{offspring}} = \beta_0 + \beta_1 \cdot \frac{Y_{\text{parent1}} + Y_{\text{parent2}}}{2} + \epsilon$$

(c) Derive the bias in $\widehat{h^2}$ when there are shared environmental effects with correlation c between relatives.

Part B: Simulation Study (15 points)

2.2 Design and implement a comprehensive simulation to study heritability estimation:

```r
# Template structure - expand this
simulate_heritability_study <- function(n_families, n_qtl, h2_true,
                                        shared_env_cor = 0, dom_variance_prop = 0) {
    # YOUR IMPLEMENTATION:
    # 1. Simulate QTL effects and frequencies
    # 2. Generate parental genotypes and phenotypes
    # 3. Simulate offspring via Mendelian inheritance
    # 4. Add environmental effects (including shared if specified)
    # 5. Estimate h2 via multiple methods:
    #    - Parent-offspring regression
    #    - Full-sib correlation (h2 = 2*r_sib for additive model)
    #    - ANOVA-based family components
    # 6. Return comparison of estimates
}
```

Requirements: - Test with different genetic architectures (oligogenic vs polygenic) - Investigate bias from shared environment and dominance - Create publication-quality plots comparing estimation methods - Discuss when each method performs best

---

Problem 3: Segregation Analysis Implementation (25 points)

Part A: Likelihood Theory (10 points)

3.1 For a binary trait following a single-locus model with genotypes AA, Aa, aa:

(a) Write the complete likelihood function for a nuclear family with n offspring, given parental phenotypes and offspring phenotype vector $\mathbf{y} = (y_1, ..., y_n)$. Include: - Penetrance parameters f_{AA}, f_{Aa}, f_{aa} - Allele frequency p - All possible parental genotype combinations

(b) Derive the EM algorithm steps for maximum likelihood estimation when parental genotypes are unknown.

(c) Show how to incorporate ascertainment bias correction for families selected because they contain at least one affected individual.

Part B: Computational Implementation (15 points)

3.2 Implement segregation analysis in R:

```r
# Template - you need to complete this
segregation_analysis <- function(family_data, model = "dominant") {
    # family_data should have columns: family_id, individual_id,
    # father_id, mother_id, affected_status

    # YOUR IMPLEMENTATION:
    # 1. Set up parameter space (p, penetrances)
    # 2. Implement likelihood calculation for each family
    # 3. Use optim() to find MLE
    # 4. Calculate LOD scores vs sporadic model
    # 5. Perform likelihood ratio tests
    # 6. Return parameter estimates, standard errors, p-values
}
```

Test your implementation with: - Simulated data from known genetic models - Real pedigree data (provide sample dataset) - Compare dominant, recessive, and additive models - Validate against published segregation analysis results

---

Problem 4: Gene Mapping Strategy and Power Analysis (20 points)

Part A: Linkage vs Association Theory (10 points)

4.1 Consider two scenarios: - **Scenario A**: $\lambda_{\text{sibling}} = 50$, $h^2 = 0.85$, significant single-locus segregation - **Scenario B**: $\lambda_{\text{sibling}} = 3$, $h^2 = 0.45$, non-significant segregation

(a) For each scenario, calculate the expected effect size (relative risk) for association studies using the relationship:

$$\text{RR} \approx 1 + \frac{2p(1-p) \cdot \beta^2}{\sigma_E^2}$$

where β is the allelic effect.

(b) Derive the sample size requirements for 80% power in linkage analysis (families needed) vs association studies (cases/controls needed) for each scenario.

(c) Create a decision framework: plot the boundary in $(\lambda_{\text{sibling}}, h^2)$ space that separates "linkage optimal" from "association optimal" regions.

Part B: Map Distance and Recombination (10 points)

4.2 Implement and compare map functions:

```r
# YOUR IMPLEMENTATION:
map_functions <- list(
    haldane = function(theta) {
        # Convert recombination fraction to genetic distance (cM)
    },
    kosambi = function(theta) {
        # Convert with interference assumption
    },
    # Add inverse functions
    haldane_inv = function(d_cM) {
        # Convert genetic distance to recombination fraction
    },
    kosambi_inv = function(d_cM) {
        # Convert genetic distance to recombination fraction
    }
)
```

Analysis Requirements: - Plot both map functions and their differences - Show how interference affects genetic distance estimates - Calculate the maximum difference between Haldane and Kosambi functions - Discuss implications for gene mapping resolution

---

Bonus Problem: Advanced Heritability Concepts (10 points)

5.1 Missing Heritability Decomposition: The lecture showed that family-based heritability estimates are often much higher than SNP-based estimates.

(a) Simulate a polygenic trait with 1000 QTLs where only 100 are genotyped ("tag SNPs"). Calculate: - True narrow-sense heritability - Heritability captured by tag SNPs only - "Missing heritability" = difference

(b) Implement a variance components model that partitions heritability into: - Common variants (MAF $>$ 5%) - Low-frequency variants (1% $<$ MAF $<$ 5%)
- Rare variants (MAF $<$ 1%) - Structural variants/CNVs

(c) Discuss how LD patterns, population structure, and gene-gene interactions contribute to missing heritability.

Submission Requirements

1. **Code**: Well-documented R functions with clear variable names and comments
2. **Plots**: Publication-quality figures with proper labels and captions
3. **Written responses**: Mathematical derivations and biological interpretations
4. **Reproducibility**: Set seeds for random simulations; provide session info

Grading Criteria: - Mathematical accuracy and completeness (40%) - Code quality and computational correctness (35%) - Biological interpretation and insights (15%) - Clarity of presentation (10%)

This assignment integrates the theoretical foundations from Lecture 02 with hands-on computational implementation, preparing you for advanced topics in statistical genetics.